Using Logistic Regression and Support Vector Classification to Predict Cancer

Bowen Zhang *

College of Engineering, Pennsylvania State University, State College, Pennsylvania, 16803, United State

* Corresponding Author Email: bbz5108@psu.edu

Abstract. This study investigates the application of machine learning (ML) algorithms in the early diagnosis of breast cancer, focusing on logistic regression and Support Vector Classification (SVC). Utilizing a dataset from Kaggle, which includes diverse clinical features from breast mass samples, the research conducts a comparative analysis of these models in terms of accuracy and interpretability. Our findings reveal that both logistic regression and SVC demonstrate high precision in distinguishing between benign and malignant tumors, with SVC showing a marginally superior performance due to its higher sensitivity and lower rate of false negatives. The study emphasizes the potential of ML in enhancing cancer diagnostic processes, highlighting the importance of non-invasive, cost-effective, and accurate diagnostic alternatives. It also addresses the challenges of model interpretability and the need for more transparent ML applications in clinical settings. This research paves the way for future advancements in medical diagnostics, offering promising directions for integrating ML algorithms into clinical decision-making and patient care.

Keywords: Logistic Regression, SVC, Prediction.

1. Introduction:

Cancer remains one of the leading causes of mortality worldwide, with millions of new cases diagnosed annually. Its impact on public health, life expectancy, and economic burden is profound, underscoring the urgent need for effective diagnostic and prognostic strategies.

The significance of accurate and timely cancer prediction cannot be overstated. Early diagnosis enhances treatment success and survival rates. However, traditional diagnostic methods, such as biopsies and imaging, can be invasive, costly, and sometimes inconclusive. In this context, the application of machine learning (ML) in medical diagnostics has emerged as a promising avenue, offering non-invasive, cost-effective, and highly accurate alternatives.

Recent research has highlighted the potential of ML algorithms to revolutionize cancer prediction and classification. For example, studies have demonstrated that logistic regression can effectively distinguish between benign and malignant tumors based on histopathological data. Similarly, Support Vector Machines (SVMs) have been employed to classify cancerous tissue with high accuracy, leveraging patterns within genetic and imaging data. These approaches have been lauded for their ability to handle complex datasets and identify non-linear relationships, which are often missed by traditional statistical methods.

However, while the existing body of research provides a compelling case for the integration of ML in oncology, there remains a gap in comparative analysis between different ML algorithms, particularly in real-world clinical settings. Moreover, the interpretability of these models and their decisions is often a concern, calling for more transparent and explainable approaches.

This paper aims to contribute to this burgeoning field by presenting a comparative analysis of two prevalent ML algorithms: logistic regression and Support Vector Classification (SVC). Utilizing a well-curated dataset, we evaluate the efficacy of these models in predicting the diagnosis of breast cancer, categorized as benign (B) or malignant (M). This research focus on the interpretability of these models, seeking to provide insights that could bridge the gap between ML predictions and clinical decision-making.

In doing so, in this research aspire to not only validate the effectiveness of these models in a vital area of public health but also to illuminate the path forward for future research that could further refine cancer prediction methodologies, ultimately aiding in the early detection and treatment of this pervasive disease.

2. Method and Data

2.1. Data:

The cornerstone of any predictive analysis is the quality and comprehensiveness of the dataset used. For this study, the cancer_dataset.csv file serves as the primary data source. This dataset, collected from website Kaggle which contains numeric dataset from diverse dimension, consists of several clinically relevant features extracted from digitized images of breast mass samples. These features include attributes like texture, radius, and perimeter, which are critical in differentiating between benign and malignant tumors.

The dataset comprises observations from hundreds of patients, each labeled with a diagnosis of 'M' for malignant or 'B' for benign. Prior to the application of machine learning algorithms, the dataset underwent rigorous preprocessing to ensure the integrity and uniformity of the data. Missing values were addressed through imputation methods, and a normalization procedure was implemented to scale the feature values, thereby eliminating the influence of variable magnitudes on the model's

2.2. Methodology

2.2.1. Logistic Regression

Logistic regression models are often created with the goal of predicting the outcomes of future patients based on each patient’s predictor variables. Regression model diagnostics measure how well models describe the underlying relationships between predictors and patient outcomes existing within the data, either the data on which the model was built or data from a different population.[1] In this study, logistic regression serves as the baseline model. The model's parameters were estimated using the Maximum Likelihood Estimation (MLE) technique. A critical step in the logistic regression analysis was the selection of features. By using techniques such as Recursive Feature Elimination (RFE), we can identify the most predictive features that contribute to the accuracy of cancer diagnosis.

2.2.2. SVC

In contrast to logistic regression, Support Vector Classification (SVC) is a more complex algorithm. It operates by creating a hyperplane or multiple hyperplanes within a high-dimensional space. These hyperplanes are instrumental for various tasks such as classification, regression, and more. The effectiveness of SVC is characterized by how well the hyperplane separates the data: ideally, it is the one that maintains the greatest distance from the nearest data points of each class, a concept commonly known as the margin.

For the purposes of this study, the SVC model was employed with a linear kernel to maintain interpretability. We experimented with different regularization parameters to prevent overfitting, ensuring that the model generalizes well to new, unseen data. The kernel trick was not used, as the primary focus was on maintaining the transparency of the model's decision-making process.

2.3. Evaluation

Model performance was assessed using a variety of metrics. The primary metric was accuracy, which provides a straightforward measure of the model's overall effectiveness. However, accuracy alone can be misleading, especially in datasets with an imbalanced class distribution. Therefore, we also computed the precision, recall, and F1-score for a more comprehensive evaluation. Then, recall is taken as the other important evaluation index here since the higher the recall is, the higher the proportion of malignant breast cancer that can be predicted.[2]

To estimate these performance metrics, we employed k-fold cross-validation. [3] This method involves partitioning the dataset into k subsets, training the model on k-1 subsets, and validating it on the remaining subset. This process was repeated k times, with each subset used exactly once as the validation data. The final performance metrics were averaged across all k iterations to produce a robust estimate of the model's predictive power.

2.4. Data Analysis

Prior to modeling, an exploratory data analysis (EDA) was conducted to gain insights into the dataset's characteristics. Histograms, box plots, and scatter plot matrices were generated to visualize the distribution of features and the relationships between them. The EDA revealed patterns and trends that informed the subsequent feature selection process and model tuning.[4]

In the feature selection phase, both univariate and multivariate methods were applied. Univariate selection methods, such as chi-square tests, were used to examine the individual strength of each feature. Multivariate methods, like logistic regression with lasso regularization, were applied to understand the collective influence of features. This dual approach ensured a balanced feature set that captures both individual and combined predictive capabilities.

The final feature set was subjected to multicollinearity checks to ensure that our models were not unduly influenced by highly correlated predictors. Variance Inflation Factor (VIF) values were calculated for each feature, with those exceeding a certain threshold removed from the dataset to mitigate the risk of multicollinearity.[5]

2.5. Model Training and Validation

The models were trained on a split of the dataset, typically using 70% of the data for training and the remaining 30% for testing. The training process involved iterative optimization of the model parameters to minimize the respective loss functions. For logistic regression, this was the log-loss, while for SVC, it was the hinge loss.

3. Result

3.1. Distribution of Diagnosis

The initial step in our analysis involved examining the distribution of diagnosis within the dataset. As shown in Fig.1, a bar chart reveals a higher frequency of benign (B) cases compared to malignant (M) ones. This imbalance in class distribution is a critical factor that can influence the performance of predictive models. Models trained on imbalanced data may exhibit a bias towards the majority class, in this case, benign tumors. It's crucial for classifiers to recognize and correctly predict malignant cases despite their lower occurrence, as these predictions carry significant implications for patient care.

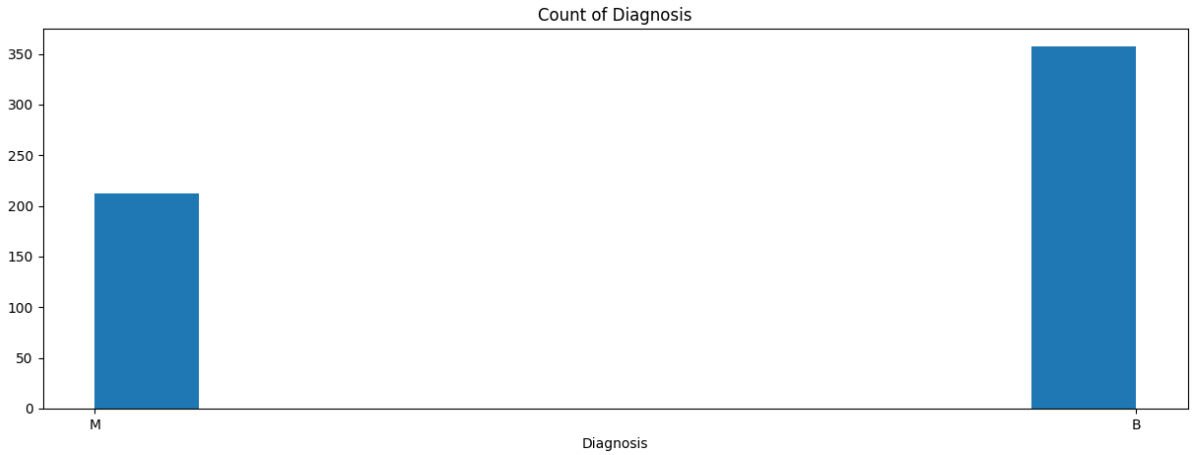

Figure 1. Count of Diagnosis

3.2. Feature Relationships and Distributions

The exploratory data analysis was furthered by examining a scatter plot matrix, as shown Fig.2. This matrix provided a visual representation of the relationships between various features such as 'radius_mean', 'texture_mean', and 'perimeter_mean'. The distribution of these features was plotted against the diagnosis categories. From the scatter plots, a noticeable separation between benign and malignant diagnoses was observed, especially when considering the 'perimeter_mean' and 'radius_mean' features. These features appear to be highly indicative of the diagnosis, as most benign cases cluster in the region of lower values, while malignant cases tend to have higher values.

The density plots along the matrix's diagonal show the distribution of individual features (Fig.2). In the context of 'radius_mean', the distribution for malignant cases skews towards higher values, contrasting with the benign cases. Such distinctions between feature distributions are promising for classification purposes as they suggest that the models can leverage these differences to differentiate between diagnoses.

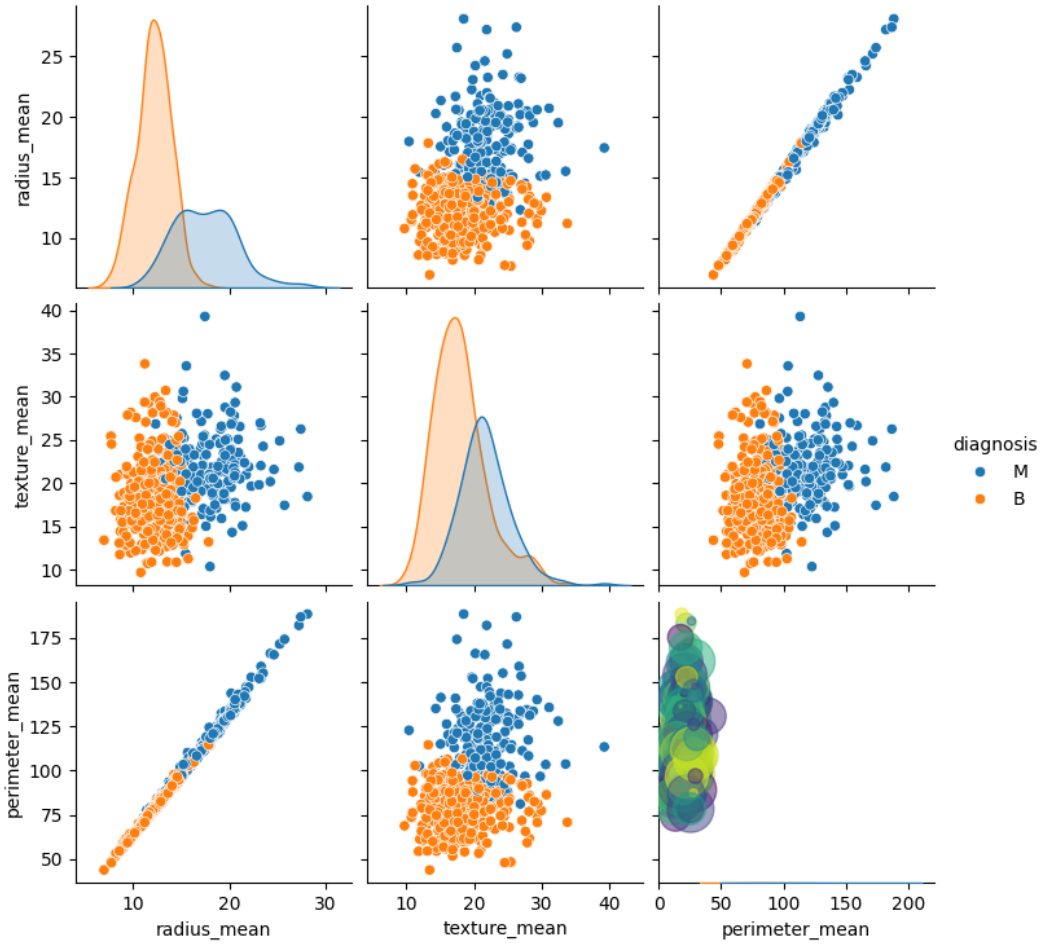

Figure 2. Combined Scatter Map

3.3. Confusion Matrices

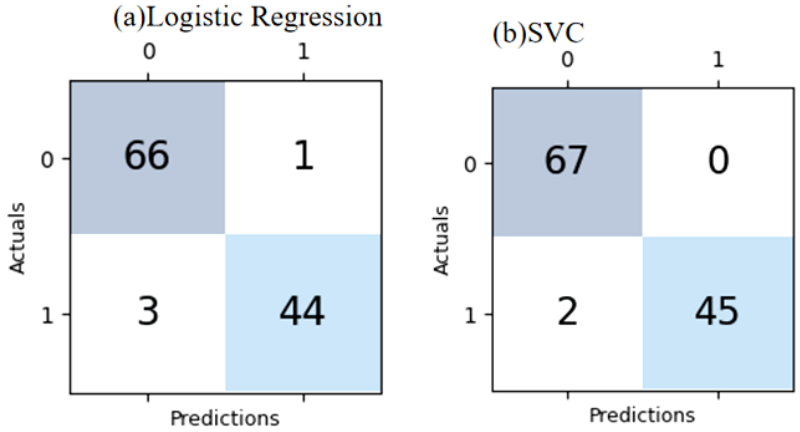

Figure 3. Matrix of Two Model

The performance of the logistic regression and SVC models was quantitatively assessed using confusion matrices, illustrated Fig.3. These matrices provide a straightforward visualization of model accuracy, depicting the number of true positives, false positives, true negatives, and false negatives.

The Support Vector Classification model exhibits a highly accurate diagnostic prediction capability, with no false positives, which is essential to avoid unnecessary treatments. It accurately identified 67 malignant cases (true positives) and correctly classified 45 benign cases (true negatives), with only 2 malignant cases misclassified as benign (false negatives), indicating a high sensitivity and an excellent negative predictive value.

The Logistic Regression model presents a strong, but slightly less accurate performance than the SVC model. It correctly identified 66 malignant cases (true positives) and 44 benign cases (true negatives), with 1 benign case misclassified as malignant (false positive) and 3 malignant cases misclassified as benign (false negatives). This shows a slightly lower sensitivity compared to the SVC model and indicates a small margin for improvement in reducing false negatives.

3.4. Precision, Recall and F1 Scores

```
Model 0
Precision is:  0.9777777777777777
Recall is:  0.9361702127659575
F1 socre is:  0.9565217391304347
Model 1
Precision is:  1
Recall is:  0.9574468085106383
F1 score is:  0.9782608695652174
```

Figure 4. Result of Two Model

In addition to the confusion matrix, the paper calculated the model's precision, recall, and F1 score to get a more nuanced view of the model's performance (Fig.4). Precision measures the accuracy of the positive predictions made by the model, while recall (also known as sensitivity) evaluates the model's ability to identify all relevant cases in the dataset. The F1 score strikes a balance between precision and recall, providing a single metric that combines both dimensions.

The logistic regression model (Model 0) achieves a high level of precision, which suggests that when it predicts that a tumor is malignant, it is correct most of the time. Recall was also impressive, indicating that the model was able to identify most malignant cases. However, the balance between precision and recall as reflected in the F1 score suggests that there is room for improvement.

Compared to logistic regression, the SVC model (Model 1) has a slightly improved recall and an impressive precision of 1.0, resulting in a higher F1 score. This improvement suggests that the SVC

model is not only accurate in its predictions but also more reliable in identifying malignant cases, which is clinically critical.

Comparing the two models, the SVC model is slightly better than the logistic regression model, especially in terms of accuracy. This suggests that the SVC model may be more appropriate for use. However, the choice between the two models may ultimately depend on the specific clinical requirements and the cost-benefit analysis of false positives versus false negatives.

In conclusion, both models show great potential for assisting in the diagnosis of cancer. They demonstrate robust performance metrics that can be integrated into the clinical decision-making process to provide a second opinion or to prescreen patients. Future work may include exploring the performance of these models on more diverse datasets, incorporating additional features or testing them in a prospective clinical setting. The ultimate goal is to improve the accuracy and reliability of cancer diagnosis, thereby improving patient outcomes and healthcare.

4. Discussion

The ability of both logistic regression and SVC models to distinguish between benign and malignant diagnoses with high accuracy carries important implications for clinical practice. It indicates the feasibility of deploying such models as assistive tools for pathologists and oncologists, potentially leading to more consistent and timely diagnoses. The use of these models could be particularly beneficial in resource-limited settings where there is a scarcity of experienced medical professionals.

Despite the promising results, several limitations must be acknowledged. First, the performance of the models is highly dependent on the quality and representativeness of the dataset. The dataset used in this study was limited in size and scope, which may restrict the generalizability of the findings. Additionally, the class imbalance present in the dataset could introduce biases that might affect the models' performance in real-world scenarios.[6]

Second, the interpretability of the models is an essential consideration. While logistic regression is inherently interpretable, the SVC model, especially with non-linear kernels, can become a 'black box', making clinical adoption challenging. Ensuring that the model's decision-making process is transparent is crucial for gaining trust from medical practitioners.

Looking ahead, there are several avenues for future research. Expanding the dataset to include a larger and more diverse population would help to validate the findings and improve the robustness of the models. Moreover, incorporating additional features, such as genetic markers or patient demographics, could enhance the models' predictive power.

Another promising direction is the integration of machine learning models with other diagnostic tools, such as imaging technologies, to create comprehensive diagnostic systems. The combination of machine learning predictions with human expertise could lead to a hybrid diagnostic approach that leverages the strengths of both.

5. Conclusion

This study applied machine learning algorithms, logistic regression, and Support Vector Classification (SVC), to the prediction of breast cancer tumors, delineating between benign and malignant classifications. A comparative analysis was conducted utilizing a dataset featuring clinically significant features from breast mass samples, sourced from Kaggle. The core aim was to evaluate the effectiveness and interpretability of these algorithms in a potential clinical diagnostic setting.

Research findings indicate that both logistic regression and SVC models exhibit high accuracy in cancer prediction, with the SVC model demonstrating a particularly strong performance characterized by an absence of false positives and a low number of false negatives. This positions the SVC model as potentially more suitable for clinical applications where the consequences of false negatives are

particularly severe. The logistic regression model, while displaying slightly lower accuracy, still offers substantial predictive capability and benefits from easy interpretability, making it a valuable asset in medical diagnostics.

Precision and recall assessments reveal that both models perform robustly, with the SVC model edging out logistic regression on these measures. The F1 scores suggest that while the models are effective, there is potential for enhancing the balance between precision and recall, especially for the logistic regression model.

Future research directions should focus on overcoming current study limitations, such as the limited scope and size of the dataset. Expanding the dataset to encompass a larger and more varied patient population, along with incorporating broader features such as genetic markers, could refine model accuracy and applicability. An intriguing avenue for future exploration lies in integrating these machine learning models with other diagnostic methods, like imaging technology, to develop multifaceted diagnostic systems.

The importance of this study is rooted in its contribution to the advancement of early and precise cancer diagnostics, which can significantly elevate patient prognosis. It lays the groundwork for the development of sophisticated tools that combine machine learning's analytical strength with the expertise of medical professionals, enhancing the quality of care in the healthcare industry.

References

[1] Meurer, William J., and J. Tolles. Logistic Regression Diagnostics: Understanding How Well a Model Predicts Outcomes. JAMA, 2017.

[2] Chen, H., et al. Classification Prediction of Breast Cancer Based on Machine Learning."Computational Intelligence and Neuroscience, 2023, 6530719 - 9.

[3] Wieczorek, J., C. Guerin, and T. McMahon. K-fold Cross-validation for Complex Sample Surveys. Stat (International Statistical Institute), 2022, 11 (1).

[4] Thivakaran, T. K., and M. Ramesh. Exploratory Data Analysis and Sales Forecasting of Bigmart Dataset Using Supervised and ANN Algorithms. Measurement. Sensors, 2022, 23, 100388.

[5] Cheng, J., et al. A Variable Selection Method Based on Mutual Information and Variance Inflation Factor. Spectrochimica Acta. Part A, Molecular and Biomolecular Spectroscopy, 2022, 268, 120652.

[6] Liu, Yuxia, et al. Influencing Factors and Prediction Methods of Radiotherapy and Chemotherapy in Patients with Lung Cancer Based on Logistic Regression Analysis. Scientific Reports, 2022, 12 (1), 21094 - 21094.